Supplementary Information: "Conservation of expression regulation throughout the animal kingdom"

Michael Kuhn, Andreas Beyer

Properties of proteins that influence the mapping

It is to be expected that properties of the considered genes have an effect on how well the genes' expression patterns can be mapped. For instance, it seems likely that genes that are well-conserved on the protein sequence level should also have conserved expression patterns. Conversely, 1:1 orthologs may appear to have dissimilar expression patterns either due to biological reasons (e.g. functional divergence) or due to technical reasons (e.g. measurement noise, inability to map the expression pattern correctly). We therefore tested eight different properties to which extent they are correlated with expression similarity (using Spearman's rank correlation coefficient). The tested properties were: number of (same-species) proteins with similar expression pattern, degree in the STRING 9.1 protein–protein interaction network (using experimental and text-mining evidence and a confidence score threshold of 0.5), number of isoforms (according to data from Ensembl, WormBase and FlyBase), number of residues, tissue specificity (Yanai et al. 2005), absolute expression level, sequence similarity between the considered proteins, and pleiotropy [for mouse proteins (Wang et al. 2010)]. Almost all properties had a significant influence (Fig. S1 and S2). The effects of sequence similarity, total expression level and degree were consistent with previous findings that these factors are inversely correlated with gene loss (Krylov 2003).

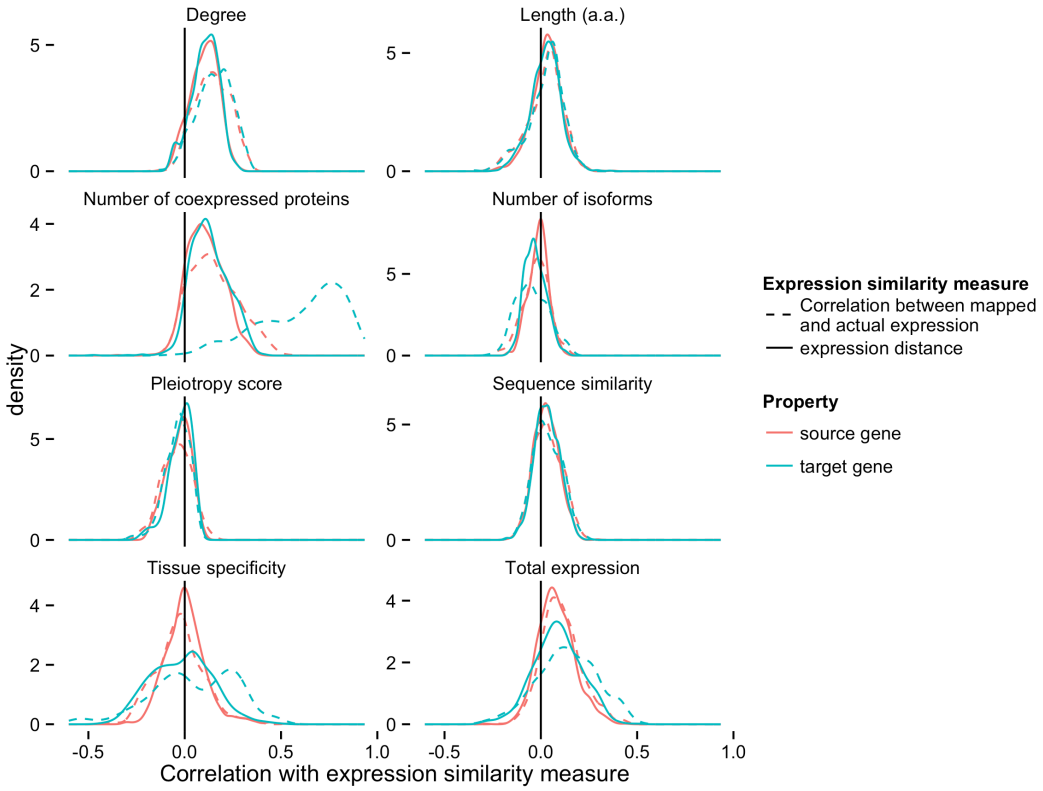

Fig. S1. Gene properties correlated with expression similarity. Different properties of the source (red) or target gene (blue) influenced the distribution of expression distances. To measure this influence, we computed the correlation between the gene properties and the expression similarity. When the correlation between mapped and actual expression patterns was used as the expression similarity (dashed line), there was a very high correlation with the number of coexpressed genes. Correcting for this (Fig. 2b, solid line), this correlation became lower. Genes corresponding to proteins with high degree (i.e. number of interactions) could be mapped better, while target genes with many isoforms resulted in a worse mapping. The influence of the factors differs with taxonomic distance, see Fig. S2.

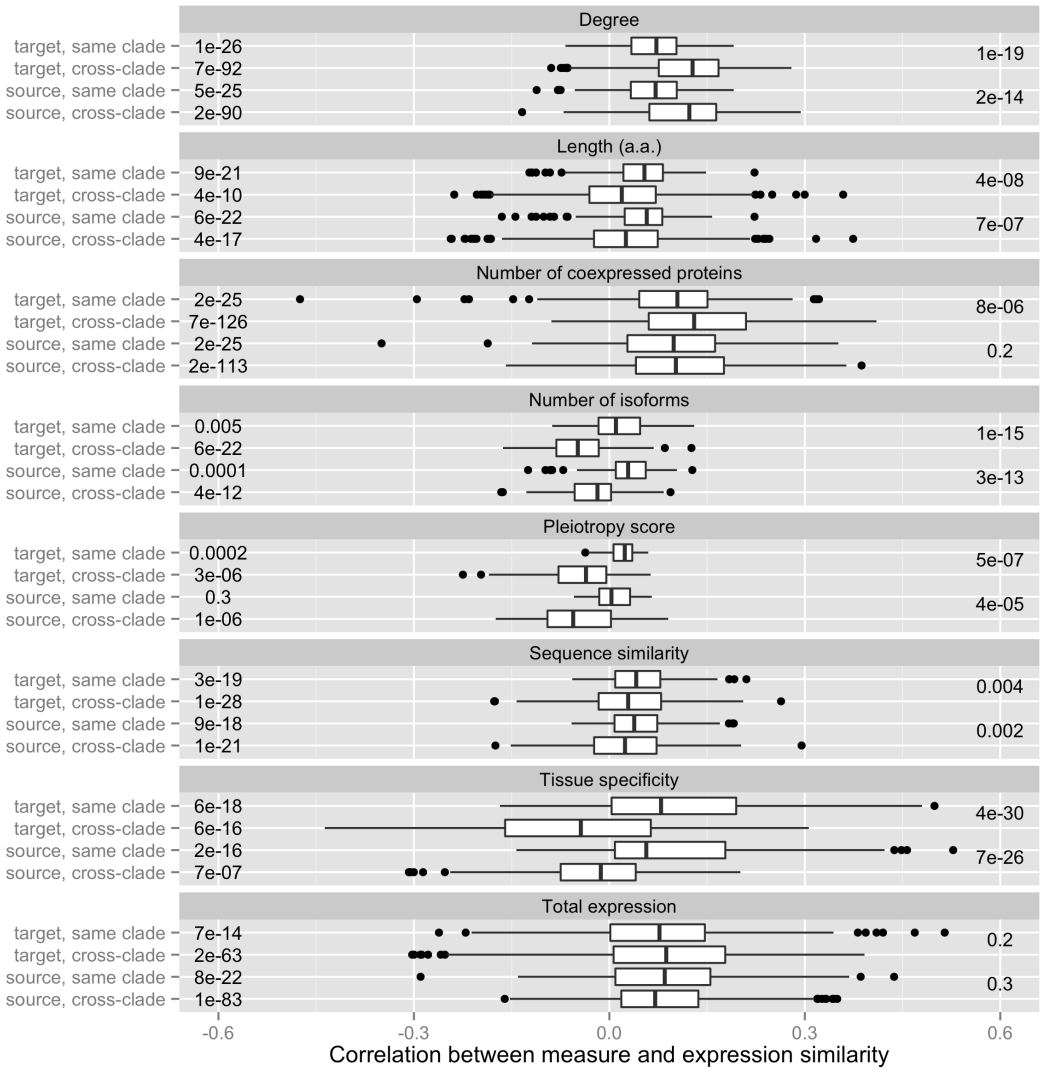

Fig. S2. Gene properties correlated with expression similarity subdivided by taxonomy. The data shown in Fig. S1 is further subdivided according to taxonomy: for some of the protein properties, the effects on the expression similarity differ strongly between dataset pairs that are in the same clade or in different clades. On the left side, p-values for a Wilcoxon signed rank test (comparing to μ=0) are shown. On the right side, a Wilcoxon rank sum test is used to test whether the effect differs between same-clade and cross-clade dataset pairs. For example, when proteins have a tissue specificity, they could be mapped better within the same clade, but worse across clades.

Benchmark 1: Identification of 1:1 orthologs

In a first benchmark, we tested whether the expression similarity could be used to identify 1:1 orthologs from top BLAST hits. For each dataset pair, we used BLAST to find the top two hits (bitscore cutoff: 100) for each protein of the source species, discarding proteins with only one hit. After training the expression mapping on an independent set of genes as outlined above, we then computed the expression similarities for the top two hits, and checked whether the gene with the lower expression distance corresponded to the actual 1:1 ortholog. For example, mapping

from fly to *C. elegans*, 67.6% of 1999 one-to-one orthologs could be correctly identified (p-value of Binomial test: 2e-57). Predictions could be ordered in different ways according to the expression distances between the two pairs of genes. For example, they could be ordered by the lowest expression distance, by the difference of the expression distances or their ratio. Of these, the difference between the expression distances performed best in distinguishing confident predictions from less confident predictions (Fig. S3). Between human and mouse (GNF dataset), 76.6% of 7304 1:1 orthologs were correctly mapped. The median fraction across all dataset pairs was 58.9%.

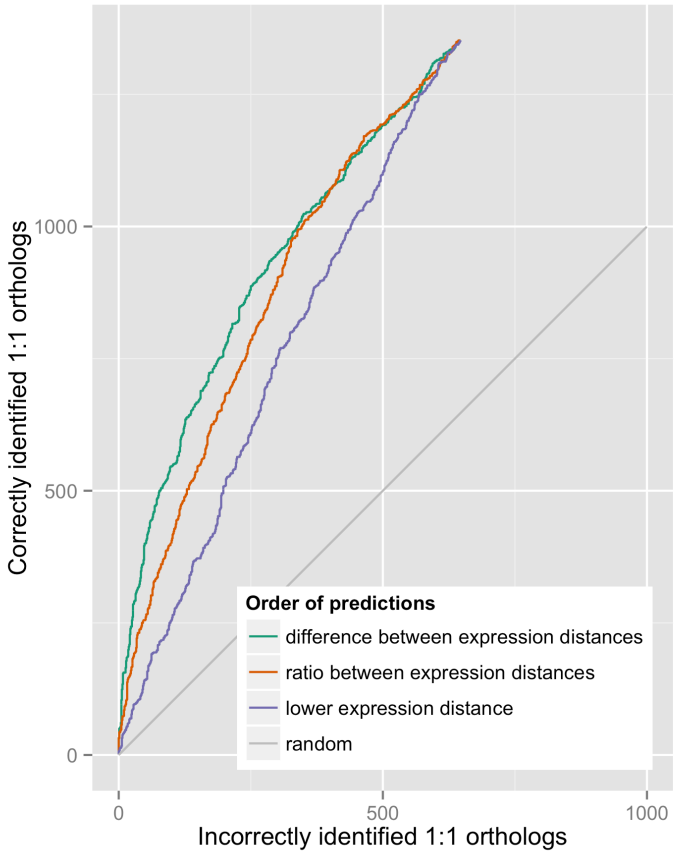

Fig. S3. Prediction of 1:1 orthologs from best hits. For each 1:1 ortholog between fly and *C. elegans*, the source gene's expression pattern is mapped to *C. elegans* and compared to the top two BLAST hits. If the mapped expression pattern is more similar to the actual ortholog, it is counted as correctly identified. Thus, a perfect prediction method would be a vertical line. Ordering the predictions by the difference between the two expression distances is the most successful strategy.

Benchmark 2: Analysis of 3D protein structure

As a further test, we checked if genes corresponding to proteins with the same structure were more likely to have lower expression distances than unrelated proteins. Using the Gene3D database (Lees et al. 2014), we determined CATH folds for all proteins that we could map to the database (resulting in 15 species and 23 datasets). For each dataset pair, we then analyzed each homologous superfamily, computing the median expression distance for all proteins of the superfamily. The superfamilies contain varying numbers of proteins, and we found a correlation between the expression distance and the size of the superfamilies (Fig. S4): Those with many members (and thus more different functions) had more diverse expression patterns. For example, mapping fly to *C. elegans*, the Spearman correlation between the number of proteins per species (using the maximum of the two species) and the median expression distance was 0.40. Between human and mouse (GNF dataset), the Spearman correlation was 0.46. Across all dataset pairs, the median Spearman correlation was 0.28.

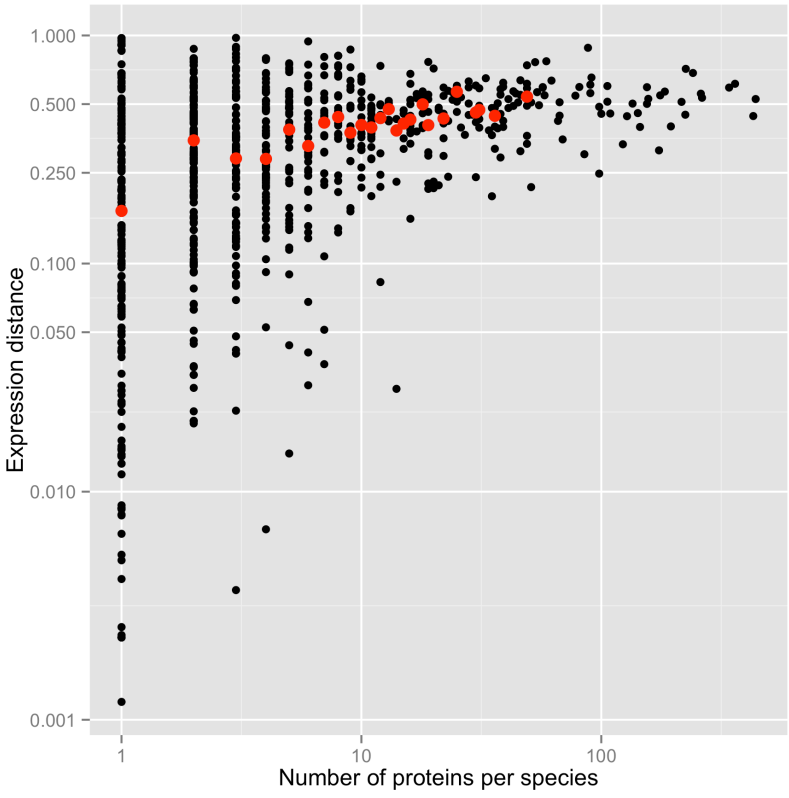

Fig. S4. Correlation between expression distance and shared protein folds. Proteins that belong to structural families with few members are more similar in their expression patterns than proteins from large families. Red dots denote the median when at least five superfamilies have the same number of proteins per species. Here, the mapping from fly to *C. elegans* is shown.

Benchmark 3: Phenologs

Finally, we used functional information to evaluate our method. We applied the phenolog concept (McGary et al. 2010) to validate that genes from different species with similar tissue expression are functionally related. Based on orthologous genes, related pairs of functional annotations (Gene Ontology terms, FlyBase and WormBase phenotypes) are predicted by looking for significant overlap between OGs that correspond to the functional annotations. For each pair of well-annotated species (mouse, human, fly, *C. elegans*), we tested all OGs excluding 1:1 orthologs. For each OG, we found the phenolog pair with the lowest p-value. For all gene–gene pairs in this OG, we then determined their expression distance and whether their functional annotation matched the phenolog pair. First, we noted that the distributions of expression distances differed between gene pairs with matched and mismatched annotations: For fly and *C. elegans*, the one-sided K-S p-value was 0.0004 and the median K-S p-value across all dataset pairs was 0.03. Second, for each OG, we looked at the gene pair with the lowest expression distance and checked if both genes matched the expected functional annotation based on the phenolog. We ordered OGs by the difference between the lowest and second lowest expression distances. Mapping fly to *C. elegans*, 50% of all top predictions had matching functional annotations, compared to an expected fraction of 43% (Fig. S5). This corresponds to a relative increase of 17% over the expected fraction. Between human and mouse, this increase is 52%. The median increase among all dataset pairs is 11%.

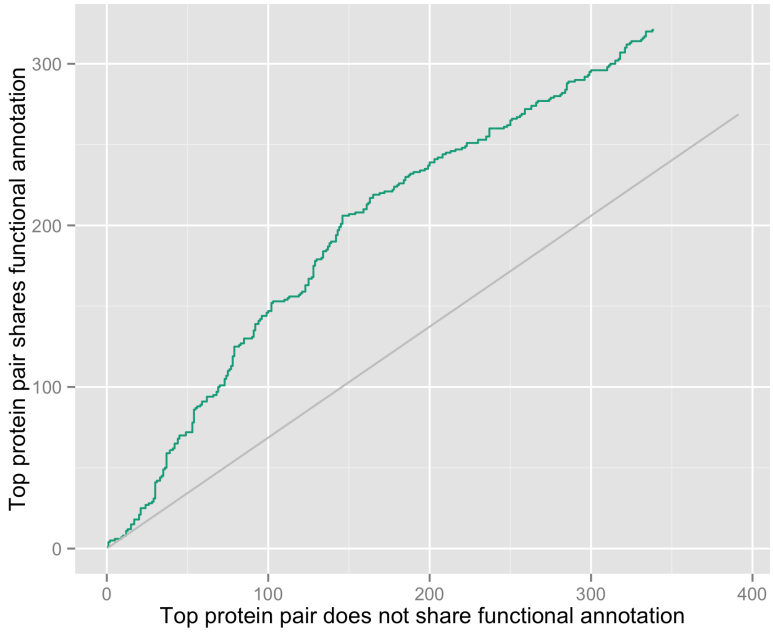

Fig. S5. Benchmarking based on phenologs. For each OG, we tested whether the gene pair with the lowest expression distance shared the functional annotation predicted by the phenolog. Predictions are ordered by the difference between the lowest and second lowest expression distance. Randomly choosing gene pairs from the OGs results in the grey line.

Conservation of tissue-specific gene expression

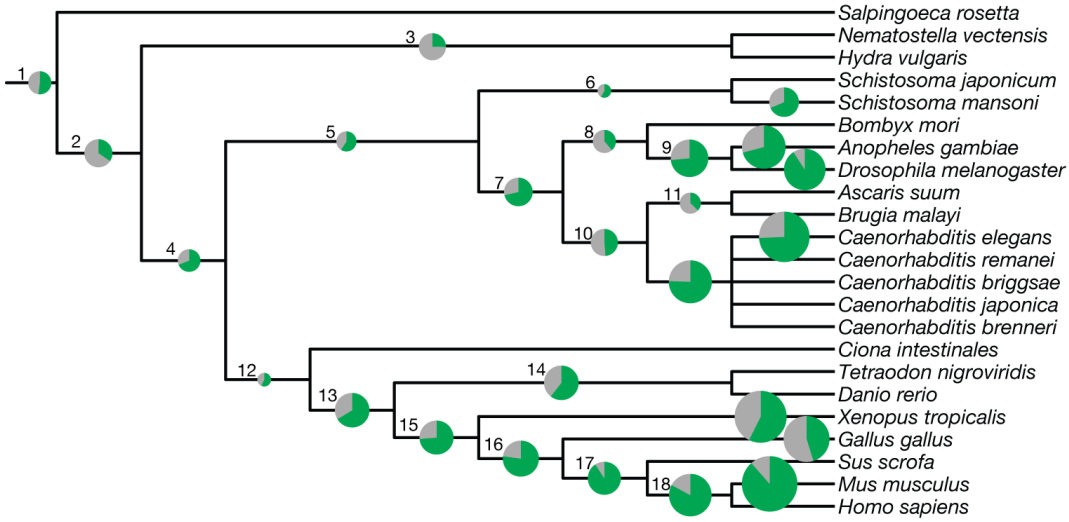

Fig. S6: Most conserved expression across animal clades. As in Fig. 6, the fraction of 1:1 orthologs with expression distances below 0.25 is shown. However, the fraction for the best dataset pair is shown instead of the median fraction for each clade. Thus, charts at species branches show how well expression patterns could be mapped using different datasets for the same species.

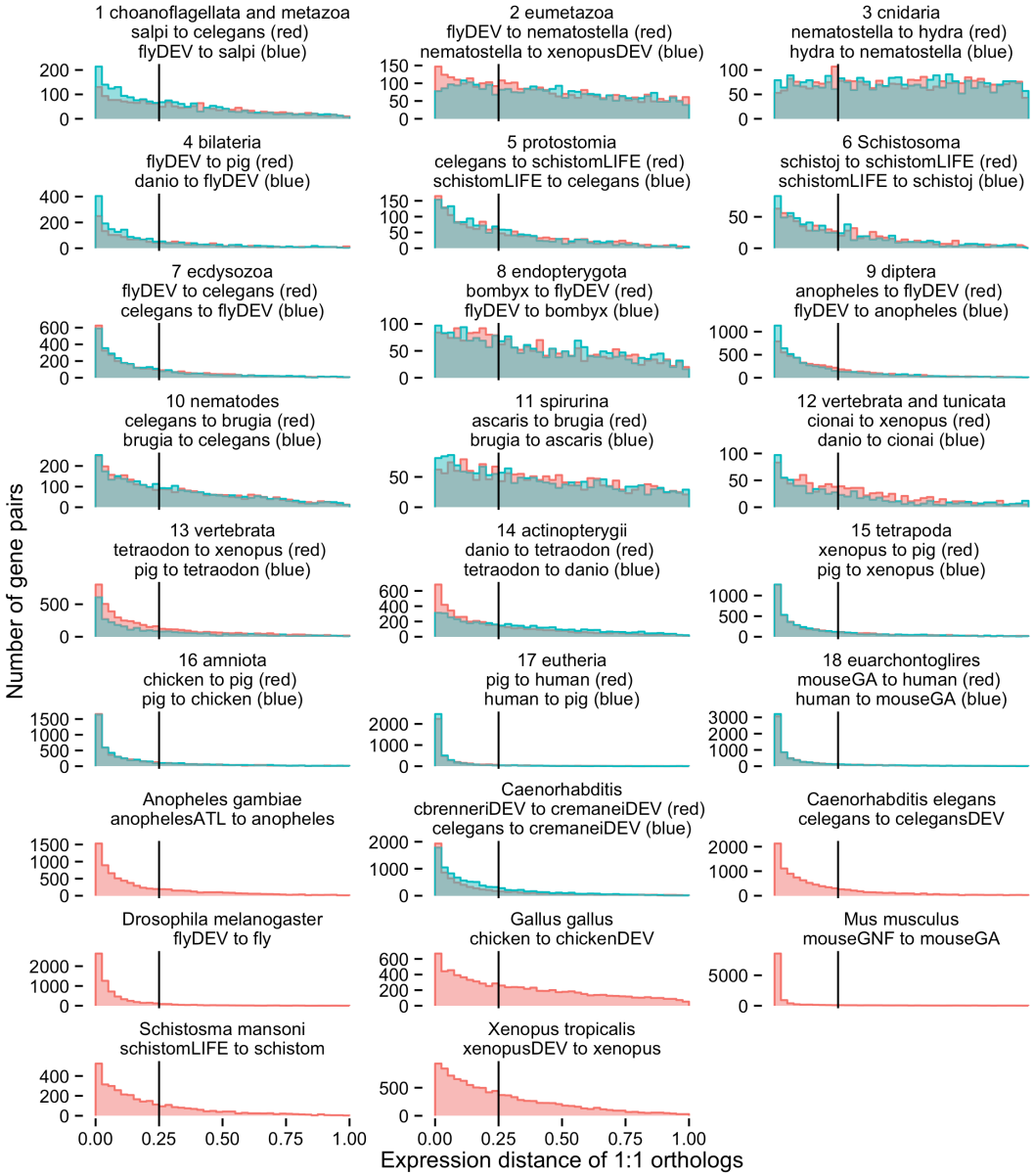

Fig. S7: Distribution of conserved expression for best dataset pairs. For each taxonomic split, the distribution of expression distances is shown for the datasets given. These datasets show the highest degree of conservation for the respective taxonomic split. See Table S1 for a description of the abbreviated dataset names.

Distribution of expression distances among non-duplicated genes

We identified sets of duplicated genes and related, unduplicated genes by constructing gene trees using GIGA (Thomas 2010). We then extracted sub-trees corresponding to a duplication event, and the corresponding related orthologous genes that emerged from speciation events. For each pair of duplication products under consideration, we considered the expression distance among the non-duplicated genes, and the expression distance across the duplication event. We then partitioned gene families based on their expression distance between the non-

duplicated genes to distinguish genes that had conserved expression patterns from those that were more variable (Fig. S9).

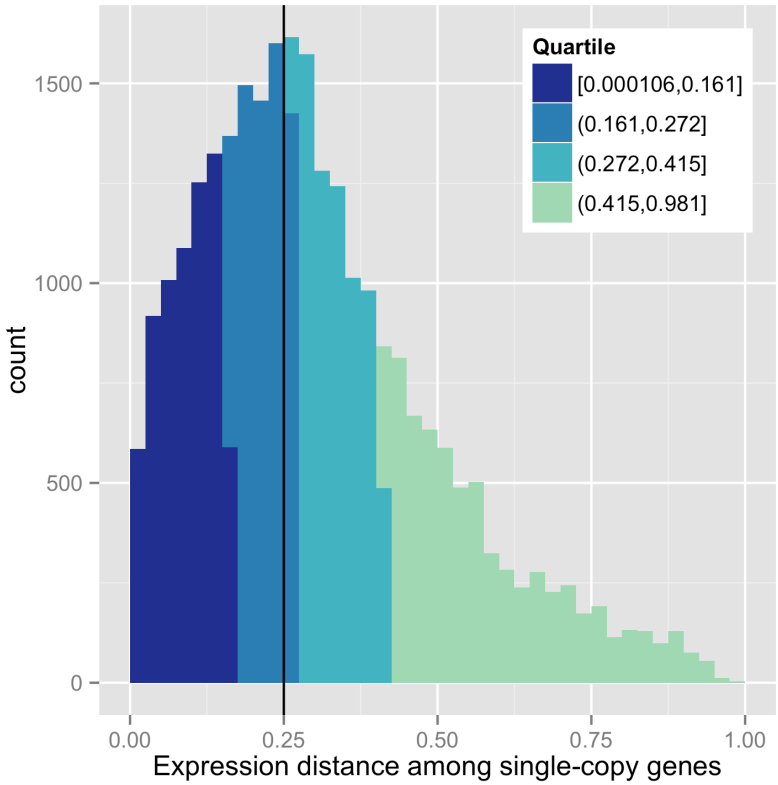

Fig. S9. Histogram of median expression distances among non-duplicated genes. As a reference for studying the fate of duplication products, we determined the median expression distance of the corresponding non-duplicated genes.

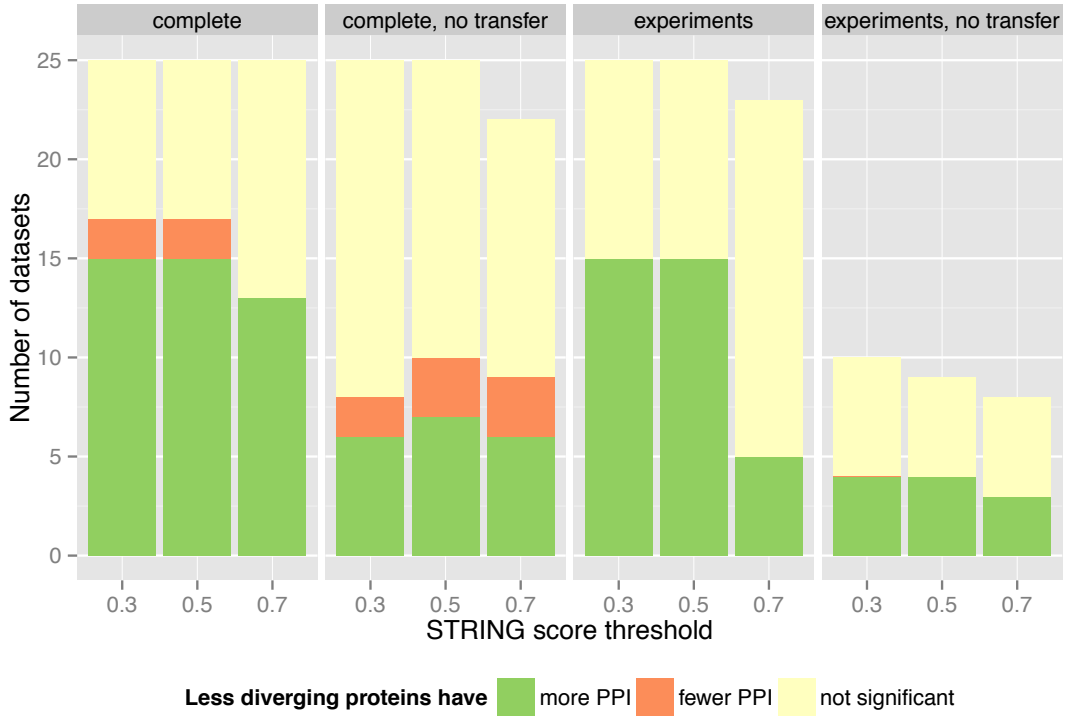

Fig. S10: Less divergent proteins had more interaction partners. For each dataset with sufficient PPI data in the STRING database, we tested which duplication product has more interaction partners using one-sided Wilcoxon signed-rank tests. In most cases, the duplication products with lower expression distances to the non-duplicated reference genes had more PPI (at a p-value cutoff of 0.05), regardless of the chosen evidence channels or score cutoffs.

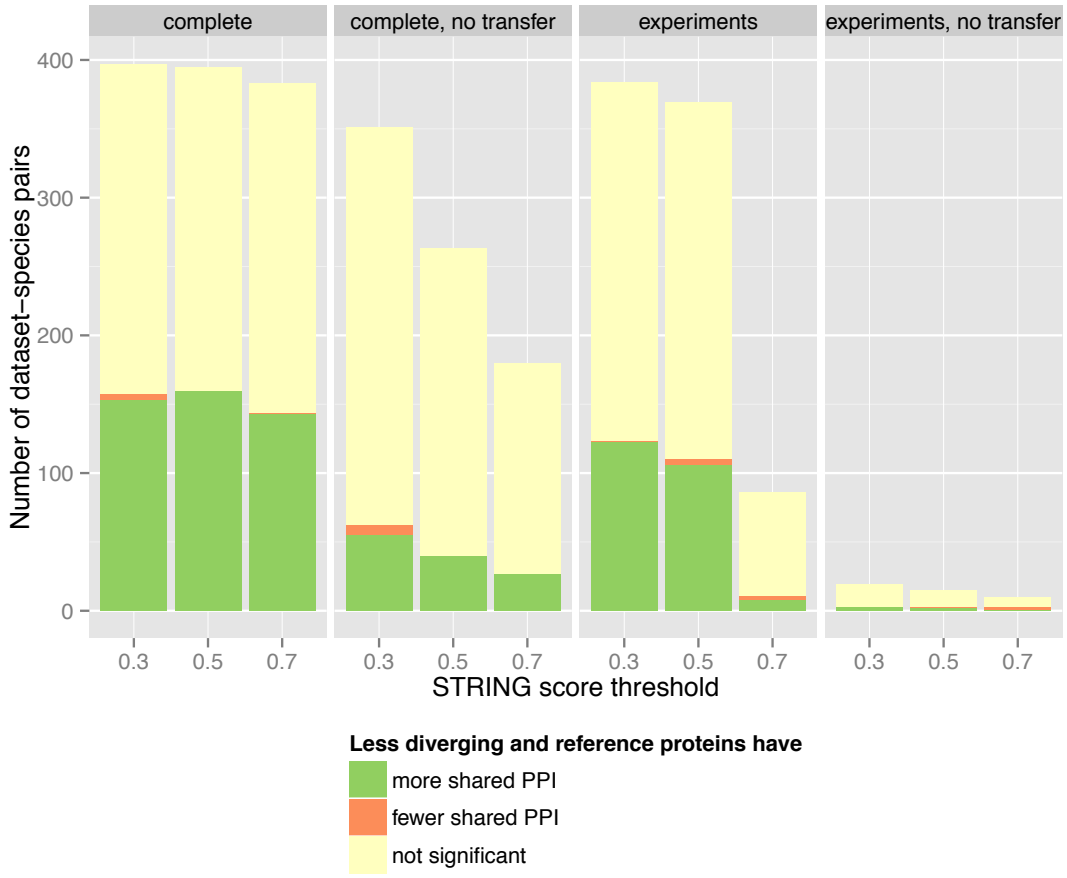

Fig. S11: Less divergent proteins shared more interaction partners with reference proteins. For each dataset-species with sufficient PPI data in the STRING database, we tested which duplication product shares more interaction partners with the non-duplicated reference protein. We compared Jaccard indices using one-sided Wilcoxon signed-rank tests and found a prevalence of duplication products with lower expression distances sharing more interaction partners with the non-duplicated reference gene.

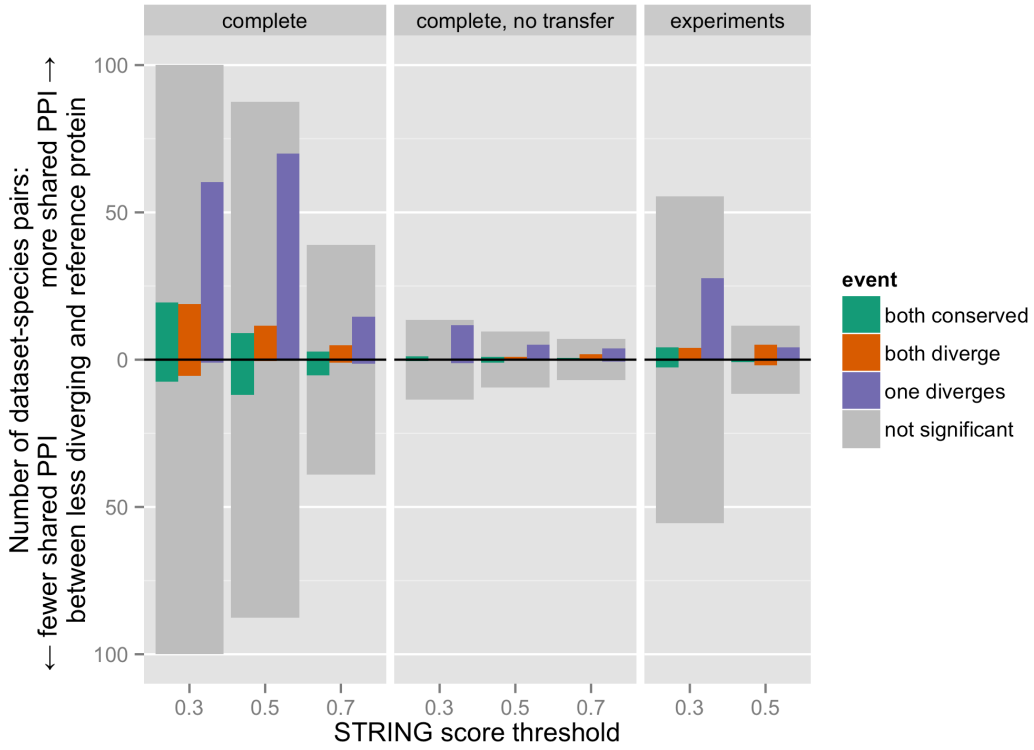

Fig. S12: Divergence of gene duplication products affects their protein–protein interactions. As in Fig. S11, we determined which duplication product shared more interaction partners with the non-duplicated reference protein. We further divided duplication events into three outcome categories (Fig. 10), and found that significant differences in the number of shared interaction partners are mainly observed for the case where one duplication product has a conserved expression pattern, while the other duplication product diverges in its expression pattern. In the other cases, i.e. when both duplication products were conserved of or both diverge, the difference in interaction partners became significant in fewer cases.

Expression divergence depends on protein function

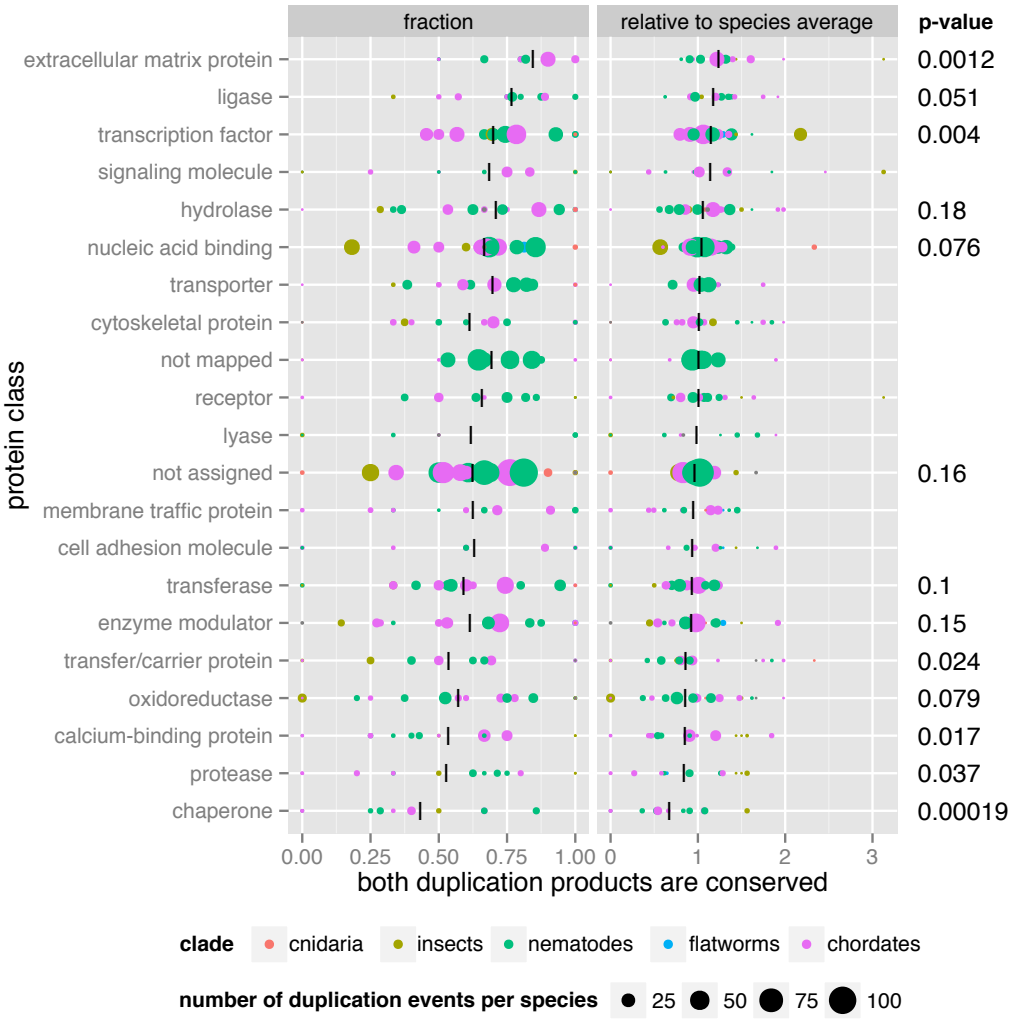

Fig. S13. Gene duplication outcomes differ between protein classes. The underlying data is the same is in Fig. 12, where OGs may be assigned to more than one protein class. Here, using an iterative procedure (see Methods), only the annotation with the most significant p-value is taken into account for each OG.

Lineage-specific expression shifts and relative expression patterns

In the main text, we investigated changing and conserved expression patterns. A previous analysis of expression patterns in six tissues across eight mammals and chicken concluded that while the expression of most genes is under purifying selection, there are also many cases of lineage-specific expression shifts (Brawand et al. 2011). However, in a re-analysis of this data, we found that these changes occurred mainly on an absolute expression level and that even across the expression shifts, the expression patterns which were reported in the original data set stayed highly correlated (Fig. S14): For the set of genes with significant

expression shifts, we found a median correlation of 0.68 between the expression patterns of the species with the expression shift and the species with unchanged expression ("outgroup"). We suspected that for some genes, the expression pattern only becomes fixed after the expression shift. Indeed, when we divided the genes into quartiles according to the median correlation within the set of proteins in outgroup, we found that in the bottom quartile the median correlation across the expression shift is 0.25, while in the top quartile the median correlation is 0.95. In other words, once an expression pattern becomes fixed, it is retained even across lineage-specific expression shifts.

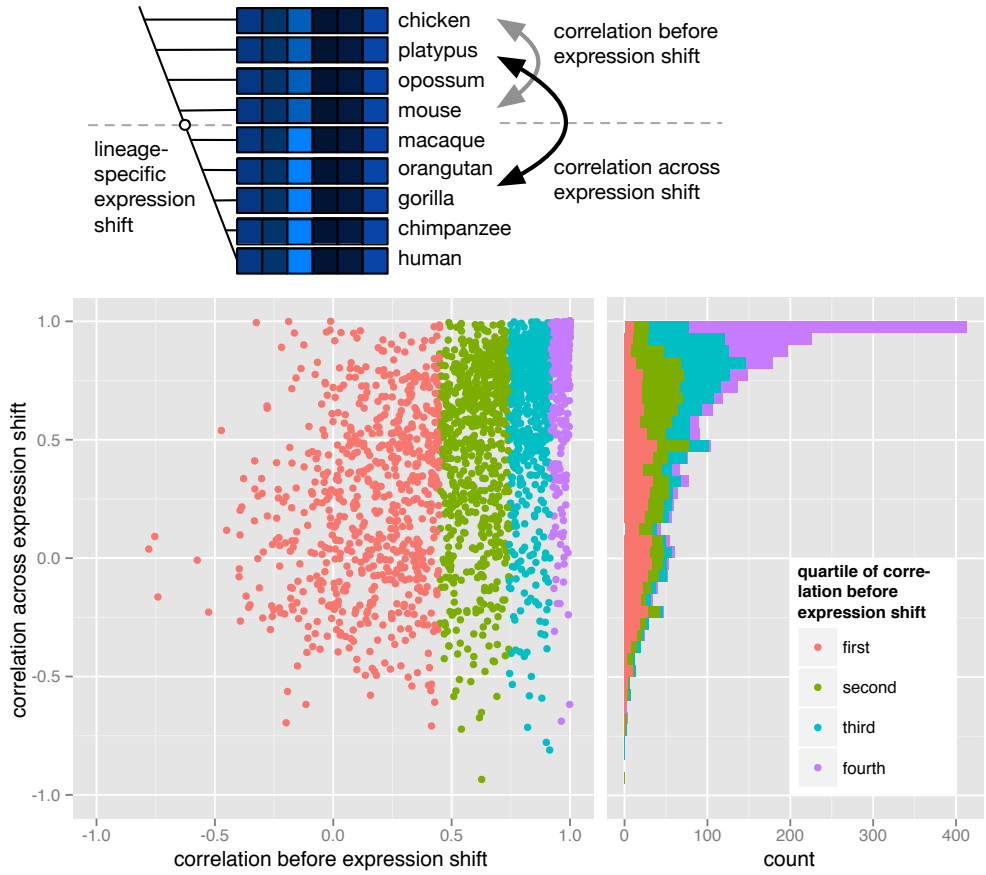

Fig. S14: Lineage-specific expression shifts do not change expression patterns. Proteins that have been reported to have lineage-specific changes in expression, e.g. between primates and non-primate species (Brawand et al. 2011) have highly correlated expression patterns even across the expression shift if the expression pattern has been fixed before (fourth quantile, purple).

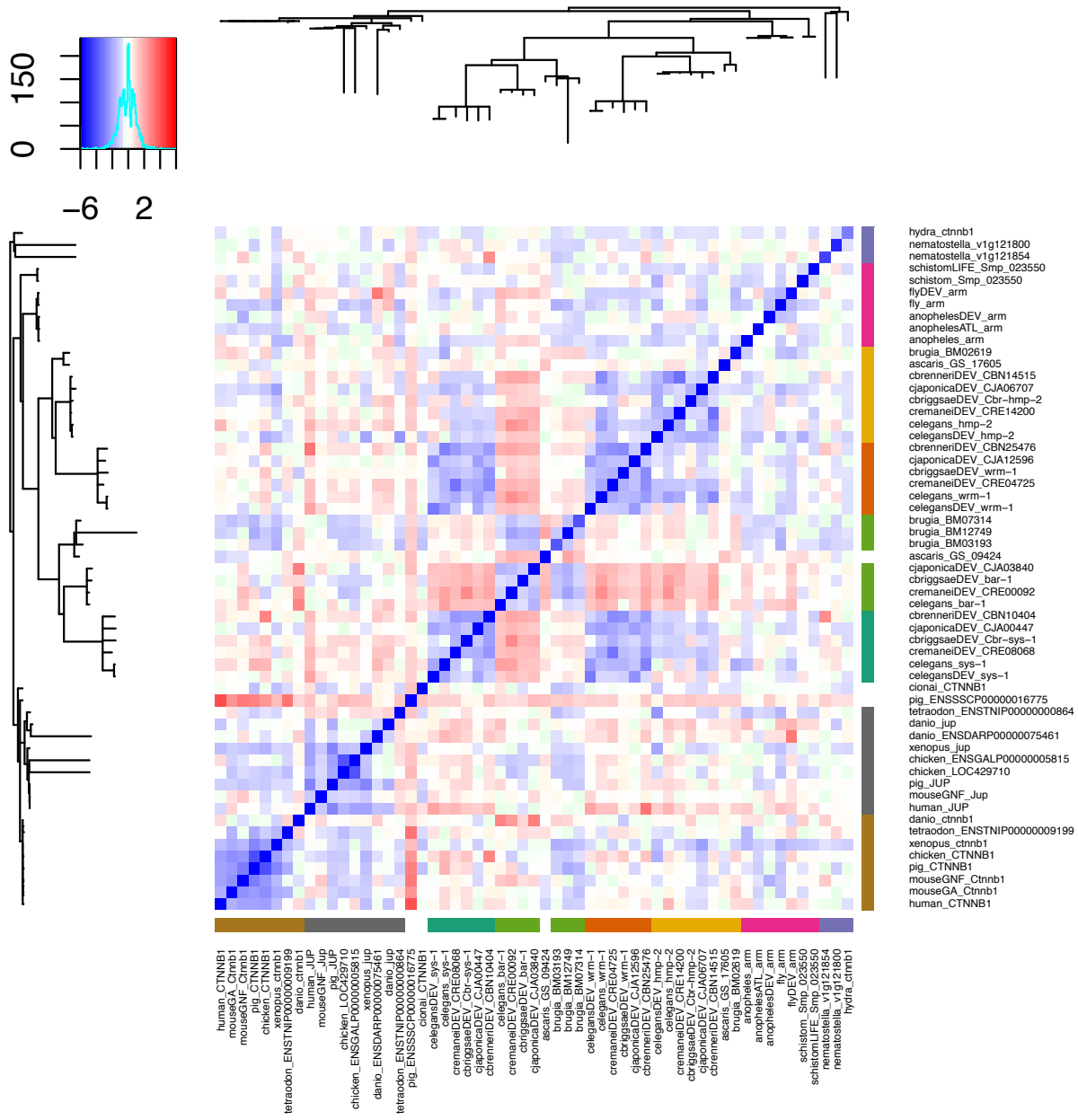

Fig. S8. Expression divergence scores of beta catenins. The heatmap shows the expression divergence scores of beta catenins throughout the animal kingdom. Significant scores are shown in blue (similarity) and red (divergence). Non-significant scores range in color from green to white to orange. For each pair of datasets, there are two expression divergence scores, depending on which dataset is the source or target dataset. Here, these two scores have been averaged. Genes are clustered according to the gene tree built from a multiple sequence alignment. The colored band along the edge of the heatmap denotes the seven groups of proteins which have been analyzed in Fig. 8. Note that one outlier, namely *Ascaris* bar-1 has been excluded due to its divergent expression pattern. An interesting feature not visible in the grouped data is that the *Brugia* bar-1 genes are more similar to the vertebrate beta catenins that the *Caenorhabditis* bar-1 genes.

References

Brawand, D. et al., 2011. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369), pp.343–348. doi:10.1038/nature10532.

Krylov, D.M., 2003. Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution. *Genome research*, 13(10), pp.2229–2235. doi:10.1101/gr.1589103.

Lees, J.G. et al., 2014. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic acids research*, 42(1), pp.D240–5. doi:10.1093/nar/gkt1205.

McGary, K.L. et al., 2010. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14), pp.6544–6549. doi:10.1073/pnas.0910200107.

Thomas, P.D., 2010. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics*, 11, p.312. doi:10.1186/1471-2105-11-312.

Wang, Z., Liao, B.-Y. & Zhang, J., 2010. Genomic patterns of pleiotropy and the evolution of complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(42), pp.18034–18039. doi:10.1073/pnas.1004666107.

Yanai, I. et al., 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics (Oxford, England)*, 21(5), pp.650–659. doi:10.1093/bioinformatics/bti042.